Towards XAI for Optimal Transport

Philip Naumann[1,2]

[1] Machine Learning Group, Technical University of Berlin, Marchstr. 23, 10587 Berlin, Germany
[2] BIFOLD – Berlin Institute for the Foundations of Learning and Data, Ernst-Reuter Platz 7, 10587 Berlin, Germany

Abstract

Transport phenomena (or distribution shifts) arise in many disciplines and are often of great scientific interest. Machine learning (ML) is increasingly used in conjunction with optimal transport (OT) to learn models for these. While XAI has improved the transparency of ML models, there has been little discussion on how to explain the factors that drive a distribution shift. Specifically, the issue of opening the OT black box has only received limited attention. Traditional classification models can distinguish between two distributions, but post-hoc explanations based on their gradients may not reveal the true reasons behind their differences. Our goal is to make OT explainable and establish XAI-OT to generate more accurate explanations for distribution shifts. We also discuss concerns regarding the accuracy of optimal transport in the presence of data issues, which we assume to have implications beyond explanations.

Keywords

Explainable AI, Optimal Transport, Distribution Shifts, Counterfactual Explanations

1. Motivation

Transport phenomena are a crucial focus of scientific research and can manifest themselves in the form of a distribution shift. Understanding these shifts can provide new insights into the factors that led to the observed changes. This can assist scientists in investigating real-world scenarios and is receiving increased attention. For examples, see the recent *DistShift* workshop [1] at NeurIPS 2022 or the *WILDS* benchmark [2].

Machine learning (ML) is popularly used to learn from data with great success. Typical tasks include classification or regression. Several methods are available to explain the classification outcome of a model (e.g. [3, 4]). They can provide valuable insights into the modeled data, helping practitioners comprehend underlying phenomena better. However, not much focus has been put on understanding distribution shifts so far [5]. Moreover, ML models themselves can be subject to these shifts causing a worsening of their performance (cf. continual learning [6]). Finding and understanding the reasons for a shift is therefore highly important. Additionally, there is evidence (see [7, 5] and section 4), that conventional classifiers that discriminate between two distributions are insufficient to accurately detect underlying shift reasons. Our work aims to fill this gap.

Various methods can be used to study the relationships between distributions. A particular framework is called *optimal transport (OT)*. Its underlying theory is well studied and comes with guarantees on the optimality of the solution (cf. [8, 9]). It solves an optimization problem that

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

p.naumann@tu-berlin.de (P. Naumann)

yields a distance between a source and target distribution—the so-called Wasserstein distance. In addition, a transportation plan with information on the allocation of mass between each source and target point is induced. This plan can be used to transport points between the two distributions. Under certain assumptions (cf. [8, 9]), the plan becomes a unique mapping function. Since the OT map is considered an 'optimal' model that represents the relationship between two distributions, it is a valuable tool for analyzing and explaining shifts [5]. We see major challenges with this, however:

It is unclear how to summarize and extract the most intrinsic and relevant information from the maps. Even though they already hold valuable information on the reasons for the shift (cf. [5]), we argue that OT does not *directly* explain the mapping in a human-comprehensible way. While it might be sufficiently transparent for a few data points in low-dimensional spaces, it quickly becomes difficult to interpret when the dimensionality increases. Because of this, we regard OT solutions as a 'black box', similar to deep neural networks (DNNs) in ML. Our goal is to move beyond this black box and make OT maps more explainable.

Furthermore, as intriguing as the theoretical guarantees of OT sound, there are also potential pitfalls where it leads to a solution that can be sub-optimal or even wrong. Even though it is an 'optimal' solution from a theoretical perspective for the data at hand, it is not guaranteed that the *data* is also optimal. Most real-world datasets are only an empirical sample of the true population. Since this is not necessarily representative, it is questionable if OT can provide a truthful approximation or even the correct solution in these cases. Statistical problems in the data are known to cause issues (e.g. [10, 11, 12] investigate the effect of outliers). We see one root cause for this in the strict mathematical formulation of OT, as it does not handle *incomplete* or *incorrect* data well. For this reason, it is especially important to consider the data and investigate it for potential issues. If we can explain OT maps, such issues may be revealed in the process and aid users in adjusting their data and model accordingly.

Apart from this, the cost function is another bottleneck for the success of the optimization. Since it is the main component of the OT objective, it heavily affects the solution. It is known that inappropriate cost functions lead to unexpected or sub-optimal solutions (e.g. [13, 14]). In the case of image data, e.g., it is usually not appropriate to apply the Euclidean distance in the input space. Still, the squared Euclidean distance is a common go-to cost function as it provides valuable theoretical properties in the context of OT (cf. [8, 9]). This suggests it is also important to carefully consider the used cost function in terms of appropriateness to the problem at hand. More expressive representations of the data might be required.

2. Related Work

Counterfactual explanations [3] can be seen as a special form of a distribution shift. These shifts occur at the decision level of a given classification model. They aim to explain the question *what would my input look like if it belonged to a different class* [3]. A typical requirement is, that the perturbation that leads to the other class should be applied with minimal effort. Additionally, the problem formulation depends on the decision function of a classifier. Without taking the nature of the data into account, it can lead to the computation of an adversarial attack [15]. Nowadays it is common that truthful counterfactual explanations should stay on the data manifold (e.g.

[16, 17]). Apart from using surrogate models, this can also be enforced as explicit constraints that guide the generation process (e.g. [18]). Some works, e.g. [19], have begun to use OT for this purpose. The main advantage over previous approaches is, that the whole distribution is considered in the process. Traditional counterfactual methods often focus on optimizing for a single instance and do not take the underlying distribution into account.

Recently, works have emerged that specifically call for a need to explain distribution shifts [20]. One particularly interesting direction uses optimal transport for this purpose [5]. The authors propose two different methods: one aims to explain shifts in a subset of features, and the other uses clustering to find differing modalities. While the former can be used to restrict the explanation to certain features, the latter can explain sub-shifts within the major shift. Both methods return a counterfactual at the data level in the form of a mean shift towards either the subset of features or the different clusters (i.e. one mean shift per cluster). Since using OT to explain distribution shifts appears to be promising, we want to investigate this direction further.

Another recent work [21] uses OT to learn a classifier whose gradient is guaranteed to point to the other class by design. This provides two interesting properties: it makes the classifier more robust to adversarial attacks and it makes the gradient more informative. Further, this property of the gradient also holds a strong resemblance to counterfactual explanations, as the authors note [21]. By following the gradient path, a potentially useful explanation emerges, instead of an adversarial example. In contrast to their work, we do not aim to learn a new classifier with OT properties but rather retrieve explanations that can be independent of a surrogate ML model.

It is known that OT maps are highly sensitive to data issues. The popular Wasserstein Generative Adversarial Networks (WGANs) [22], for example, were proposed as a more robust alternative to standard GANs [23]. They use an OT-based loss function to learn the generative model. Since OT also considers the geometry of the data, the authors found this loss design to be more robust to the issue of mode collapse [22]. However, in [10] the authors found that WGANs are still affected by other issues. They are not robust to outliers in the data which can lead to undesired image generations. This can be a serious practical issue, as there is no guarantee that the model will not produce inappropriate images. Moreover, it was shown in [24] that WGANs are not necessarily learning the correct Wasserstein distance, even though they specifically optimize for it. Surprisingly, they still perform well on their main task of data generation. This raises the question of how important an ‘optimal’ transport is.

Recently, other transport-based models like Cycle-GAN [25] have been investigated in terms of data issues as well. In [14], the authors criticize that the mappings of Cycle-GAN are seemingly random. They improve this by incorporating an OT loss to consider the geometry of the data and produce more coherent mappings. Moreover, they show that Cycle-GAN transport can fail to align with human expectations in the presence of missing data. This indicates that data issues are a concern for other transport-based models as well, giving the topic of detecting such problems relevance beyond OT.

3. Research Questions and Approach

While the black box of classical machine learning models like classifiers has been successfully opened (cf. [26]), explanation techniques for models of distribution shifts have only received little attention. Recently, optimal transport has been used to explain distribution shifts [5]. However, we argue that OT models are still largely a black box as they are not directly human-comprehensible. We aim to fill this gap by investigating two primary topics to establish XAI-OT:

(1) Can we design XAI techniques to faithfully explain OT models so that they become interpretable for humans? We want to develop XAI methods for opening the OT black box. Our investigation will assess whether existing XAI techniques apply to distribution shifts, or if specific techniques, building tightly on OT maps, need to be designed. The preliminary evidence in section 4 suggests that the gradient of classifier DNNs is not suitable for this task in some cases and that OT provides a more truthful explanation. In practice, this may take the form of attributing the Wasserstein distance across input features, either globally or at the level of individual data points. For this purpose, we will investigate perturbation methods, e.g. gradient-based, or propagation-based techniques like layer-wise relevance propagation (LRP) [4]. Notably, exploring the Kantorovich dual representation of OT (e.g. [27]) appears to be promising for this, since it can be expressed as a function of the input. Additionally, we will evaluate the faithfulness and interpretability of the generated explanations. Toward this end, we will explore techniques such as pixel-flipping or human evaluations.

(2) Can we use XAI-OT to gain insights into real-world transport phenomena? As the consideration of OT for explaining distribution shifts shows promise [5], we want to further investigate its potential. Concretely, we aim to use XAI-OT to explain real-world transport phenomena, like simulated processes or shifts between different data sources. XAI-OT may also be used to inspect the quality of the OT model itself, in particular, diagnosing potential issues such as overfitting effects or the reliance on spurious correlations in the data (cf. [28]). This way, it can help to find out why a mapping failed to meet expectations, so a user can act upon it and correct the model or data. We will also explore the intriguing connection to counterfactual explanations, as highlighted in, e.g., [5, 21, 19]. Our goal is to understand how effective OT is for generating explanations and in which contexts it is most beneficial. Finally, we aim to investigate its usefulness for uncovering novel relationships across various domains, particularly in fields of significance such as medicine or chemistry.

4. Preliminary Results

We now discuss our preliminary analysis, suggesting that existing XAI techniques may not be amenable for explaining distribution shifts and that specific XAI solutions for OT need to be developed. In fig. 1 we demonstrate the divergence between the classifier and OT gradient. The *target* data represents a data shift of the *source* data that only occurred on the x-axis. This means only one feature is relevant to explain the shift. A classifier $f : \mathbf{X} \to \{0, 1\}$ was trained

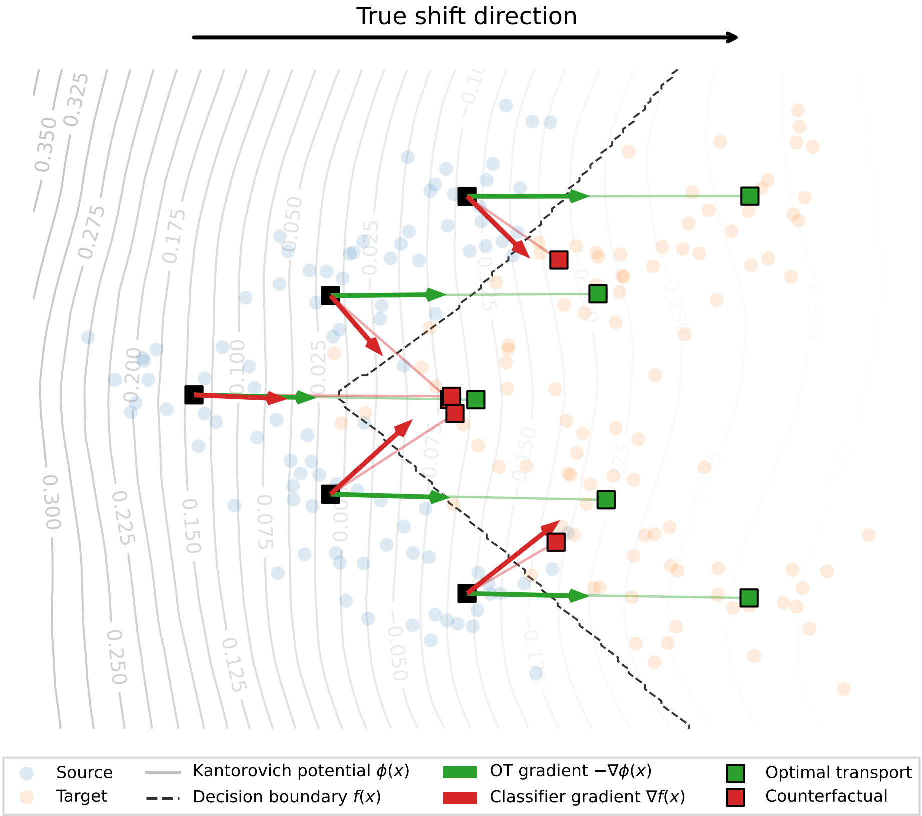

Figure 1: A comparison of classifier vs. OT explanations in the context of distribution shifts.

to discriminate between the two datasets. Additionally, $\phi : \mathbf{X} \to \mathbb{R}$ is the so-called Kantorovich potential [9] function that was learned by a different neural network.

Feature relevance: gradient vs. OT: Even though the decision boundary of f is well learned to discriminate between the two classes (i.e. the dashed line between source and target in fig. 1), the gradients do not explain the data shift correctly. As expected, they point to the decision boundary and suggest that the y-axis is also relevant for the shift. Such false attributions of feature relevance are a concern in neuroscience [7], giving this issue important practical implications. The OT potential, on the other hand, detects the true shift cause. The contour lines of the potential function are depicted in solid and are approximately orthogonal to the true shift direction. This behavior of the potential was also used in [21] to learn classifiers whose gradients are aligned with the distributions.

To conclude, this simple example illustrates why the gradient of a classification model can be deceptive as an explanation for distribution shifts. It does not account for the underlying data distribution and gives too much weight to uninvolved features. Subsequent XAI techniques that make use of the gradient information are therefore expected to provide a wrong explanation for the occurrence of the shift.

Counterfactual explanations: Another interesting observation can be made in terms of counterfactual explanations. The red squares in fig. 1 exemplify simple counterfactuals that were computed to possess high target class confidence ($95\% \leq$) according to the classifier. As can be seen, they are on the data manifold and admit to the shortest perturbation criterion. However, when we compare them to the OT locations (green squares), it becomes obvious that just staying on the manifold is not necessarily sufficient. The original, relative representation of the source points within their distribution is not reflected well in the target distribution in the case of the classifier counterfactuals. In contrast, the OT map provides better target representations as it considers the whole distribution. Moreover, simple counterfactual explanations likely have difficulties in reaching the outer points that the OT map hits. Some parts of the distribution could be hardly reachable for a standard counterfactual. We think that exactly this benefit of OT is crucial for truthful explanations.

Besides, even though the previous examples suggest that OT is an intriguing tool for explaining data shifts, it is unclear how to summarize the map. Moreover, OT does not always work well as data issues can distract the map. For these reasons, we want to focus our research in the direction of XAI-OT.

5. Outlook

Finding the true factors that drive data shifts is valuable information. Gaining such knowledge has wide-ranging implications in other scientific fields. Thus, we aim to leverage XAI for optimal transport. A major goal is to propose a method that can uncover previously unknown relationships, possibly helping scientific research in significant fields such as medicine.

Optimal transport is increasingly used in various fields of ML. We assume that many users do not pay specific attention to the impact of data quality or the utilized cost function on OT. It might even be a mostly unknown pitfall since OT losses may still appear to work in practice. Thus, we want to raise awareness of these issues and their possible consequences on OT. More robustness will likely lead to even better results. This could mean, e.g., having a human-in-the-loop type of feedback. That is, a user may post-hoc diagnose their OT model with the tools we provide and possibly act to resolve any revealed issues.

Lastly, there is evidence that our hypotheses on the statistical data issues do not only apply to optimal transport, but to other transport-based models (e.g. Cycle-GAN) as well. For example, [14] shows that Cycle-GANs cannot naturally handle data gaps, which leads to wrong mappings. In a broader scope, data issues are already known to cause problems in classical ML models [28]. This means, our investigations aim to extend the literature in this direction by analyzing the behavior and robustness of transport-based models in general.

Acknowledgments

I would like to express my gratitude to Grégoire Montavon and Jacob Kauffmann for their invaluable assistance with this project. We gratefully acknowledge funding from the German Federal Ministry of Education and Research under the grant BIFOLD24B.

References

[1] NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications, https://sites.google.com/view/distshift2022, 2022. Online; accessed 15-April-2024.

[2] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, et al., WILDS: A Benchmark of in-the-Wild Distribution Shifts, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, pp. 5637–5664.

[3] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harvard Journal of Law and Technology 31 (2018) 841–887.

[4] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller, Layer-Wise Relevance Propagation: An Overview, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer International Publishing, Cham, 2019, pp. 193–209. doi:10.1007/978-3-030-28954-6_10.

[5] S. Kulinski, D. I. Inouye, Towards Explaining Distribution Shifts, in: Proceedings of the 40th International Conference on Machine Learning, PMLR, 2023, pp. 17931–17952.

[6] Z. Cai, O. Sener, V. Koltun, Online continual learning with natural distribution shifts: An empirical study with visual data, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8281–8290.

[7] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, et al., On the interpretation of weight vectors of linear models in multivariate neuroimaging, NeuroImage 87 (2014) 96–110. doi:10.1016/j.neuroimage.2013.10.067.

[8] C. Villani, Optimal Transport: Old and New, Grundlehren Der Mathematischen Wissenschaften, Springer Berlin Heidelberg, 2008. doi:10.1007/978-3-540-71050-9.

[9] G. Peyré, M. Cuturi, Computational Optimal Transport, 2020. doi:10.48550/arXiv.1803.00567.

[10] Y. Balaji, R. Chellappa, S. Feizi, Robust Optimal Transport with Applications in Generative Modeling and Domain Adaptation, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 12934–12944.

[11] D. Mukherjee, A. Guha, J. M. Solomon, Y. Sun, M. Yurochkin, Outlier-robust optimal transport, in: Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021-07-18/2021-07-24, pp. 7850–7860.

[12] S. Nietert, Z. Goldfeld, R. Cummings, Outlier-robust optimal transport: Duality, structure, and statistical analysis, in: International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event, volume 151 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 11691–11719.

[13] C.-H. Lin, M. Azabou, E. Dyer, Making transport more robust and interpretable by moving data through a small number of anchor points, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, pp. 6631–6641.

[14] E. de Bézenac, I. Ayed, P. Gallinari, CycleGAN Through the Lens of (Dynamical) Optimal Transport, in: Machine Learning and Knowledge Discovery in Databases. Research Track, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2021, pp. 132–147. doi:10.1007/978-3-030-86520-7_9.

[15] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, et al., Intriguing properties of neural networks, 2014. doi:10.48550/arXiv.1312.6199.

[16] M. Pawelczyk, K. Broelemann, G. Kasneci, Learning Model-Agnostic Counterfactual Explanations for Tabular Data, in: Proceedings of The Web Conference 2020, WWW '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 3126–3132. doi:10.1145/3366423.3380087.

[17] A.-K. Dombrowski, J. E. Gerken, K.-R. Müller, P. Kessel, Diffeomorphic Counterfactuals With Generative Models, IEEE Transactions on Pattern Analysis and Machine Intelligence 46 (2024) 3257–3274. doi:10.1109/TPAMI.2023.3339980.

[18] P. Naumann, E. Ntoutsi, Consequence-Aware Sequential Counterfactual Generation, in: Machine Learning and Knowledge Discovery in Databases. Research Track, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2021, pp. 682–698. doi:10.1007/978-3-030-86520-7_42.

[19] L. You, L. Cao, M. Nilsson, DISCOUNT: Distributional Counterfactual Explanation With Optimal Transport, 2024. doi:10.48550/arXiv.2401.13112.

[20] J. Liu, T. Wang, P. Cui, H. Namkoong, On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets, in: Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2023.

[21] M. Serrurier, F. Mamalet, T. Fel, L. Béthune, T. Boissin, On the explainable properties of 1-Lipschitz Neural Networks: An Optimal Transport Perspective, in: Thirty-Seventh Conference on Neural Information Processing Systems, 2023.

[22] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein GAN, 2017. doi:10.48550/arXiv.1701.07875.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., Generative Adversarial Nets, in: Advances in Neural Information Processing Systems, volume 27, Curran Associates, Inc., 2014.

[24] A. Korotin, A. Kolesov, E. Burnaev, Kantorovich strikes back! Wasserstein GANs are not optimal transport?, in: Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 13933–13946.

[25] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2242–2251. doi:10.1109/ICCV.2017.244.

[26] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, K.-R. Müller, Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications, Proceedings of the IEEE 109 (2021) 247–278. doi:10.1109/JPROC.2021.3060483.

[27] A. Makkuva, A. Taghvaei, S. Oh, J. Lee, Optimal transport mapping via input convex neural networks, in: Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020-07-13/2020-07-18, pp. 6672–6681.

[28] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking Clever Hans predictors and assessing what machines really learn, Nature Communications 10 (2019) 1096. doi:10.1038/s41467-019-08987-4.